Extended abtract

EXTENDED ABSTRACT

Title: Assessing the impact of cultural and creative industries on the wealth of the regions through causal Machine Learning

[Preliminary draft – please do not circulate or cite without permission of the author]

Authors and e-mails:

Rafael Boix Domenech (rafael.boix@uv.es)[1,3]
Pau Rausell Köster (pau.rausell@uv.es) [2,3]
Jordi Sanjuan Belda (jordi.sanjuan@uv.es) [2,3]

Department: (1) Departament d'Estructura Econòmica; (2) Departament d'Economia Aplicada ; (3) Econcult

University: Universitat de València

Subject area: donde os venga mejor entre: S02 – Big Data y Algoritmos de Machine Learning en Ciencia Regional; 2. Economía del conocimiento, creatividad y geografía de la innovación; 1. Crecimiento, convergencia y desarrollo regional y urbano.

Abstract: This paper compares the causal impact of cultural and creative industries (CCIs) on the productivity of the European regions. Previous works have managed to show that the causal effects of CCIs on productivity and per capita GDP of regions, countries and municipalities are positive and high on average. This work introduces recent techniques of Causal Machine Learning are applied and compared to other traditional procedures. The results are particularly relevant for the economic policy since they allow to confirm two previous results of the literature: that the average impact of the CCIs is positive for most of the regions (although not in all) and that also this impact is high on average.

Keywords: *cultural and creative industries; European regions; Machine Learning*
JEL codes: R11; R12; R58

1. Introduction

The objective of this paper is to explore the application of machine learning methods (ML) to calculate the impacts of the cultural and creative industries (CCIs) on the income per capita and the labour productivity of places.

The model is applied to two data samples: the first from 78 countries on 5 continents, and the second from 275 European NUTS 2 regions.

For the estimation, a wide range of ML methods is used, which makes it possible to compare the characteristics of the results of each one and its predictive capacity, as well as comparing between traditional and newer methods. Among the traditional ML methods, we used: Ordinary Least Squares (OLS), Instrumental Variables (IV), and fixed effects (FE). New ML methods include: polynomial models/Generalized Additive Models (GAMs), local linear least squares (LLLS), regression trees, random forests (RF), causal forest, gradient boosting (XgBoost), and deep learning neural networks (NN) based on backpropagation and transformers.

This article contributes to the literature by providing evidence on the use of ML methods in estimating the effects of creative industries within the framework of regression and prediction. This contribution is based not only on quantitative results but also on qualitative assessments of the characteristics of each family of methods and when they can be more advantageous.

The article is divided in five sections. After the introduction, section two introduces the logic in the selection of ML methods and explains each method. Section three explains the model and data. Section four describes the results. Finally, section five discusses the conclusions.

2. Machine learning for calculating and analyzing the effects of creative industries

The application of artificial intelligence and machine learning to the creative industries is increasingly evident (Lazzeretti, 2021; Anantrasirichai & Bull, 2021). However, there is no precedent for analyzing the impacts of creative industries from a data science and ML perspective.

In this paper we propose an orderly approach to the use of ML to calculate the impacts of CCIs in the most common case: that of regression with a continuous dependent variable and an unlimited number of explanatory variables. The discussion also covers the case in which the explanatory variable of the impact of the CCIs is endogenous, although as the one in empirical application this variable is continuous, matching models will not be discussed.

That covers from the range of the most basic and traditional models ML to deep learning models. This range encompasses four families: parametric models, non-parametric models, trees and ensembles, and neural networks.

It is difficult to cover all the existing models, so within each range we will focus on the most used approaches, encompassing 14 different models. The selection of models is applicable to discrete dependent variables, except in the case of parametric models, where it would be enough to replace the selection by its discrete equivalents (e.g. logit, etc.).

The selected parametric linear models are ordinary least squares (OLS), fixed effects (FE) and Instrumental Variables (IV). The nonparametric models are Generalized Additive Models (GAM) and Local Linear Least Squares (LLLS). The third family encompasses Conditional Regression Trees (CRT) and three types of trees' ensembles: Random Forests (RF), Causal Random Forests (CRF), and Extreme Gradient Boosting (XgBoost). Deep learning includes two types of neural networks based on backpropagation algorithms and neural network based on transformers optimized for tabular data.

3. Model and data

3.1. Model

The article departs from a framework derived from the endogenous growth theory proposed by Soler & Boix (2017). In this model, CCIs increase product variety of the economies and this variety spread across a larger number of activities increasing the productivity and the Gross Domestic Product (GDP) per capita (Boix and Soler, 2017). The framework is synthesized in a log-linear equation in which GDP per capita depends on the participation of creative industries in the economy and on other usual factors in

growth models (capital stock, technical progress, etc.) (De Miguel & Rausell, 2021). This equation has been estimated in other previous papers (Peiró & Rausell, 2021; Boix, De Miguel & Rausell, 2021), which facilitates the comparison and discussion with previous applications.

$$\ln\left(\frac{GDP}{P}\right) = b\,ln\delta + b\lambda\,lns_R + lns_y + a\ln s_k - b\,lng_A + b\lambda\,lnL - a\ln(n + g_A + d) + \ln\left(\frac{L}{P}\right)$$

,in which the GDP per capita (GDP divided by population *P*) for a year *t* depends on the rate at which new ideas are created (δ), the share of persons employed in the creative sector (s_R), the share of persons employed in the rest of the economy (s_Y), the intensity of capital per worker (s_K), g_A represents the growth rate of the ideas, the population growth rate (*n*), the rate of depreciation of capital (*d*), and the ratio of labour to population (L/P). In the equation, λ measures the existence of scale economies, $a = \propto/(1-\propto)$, and $b = 1/(1-\phi)$, where α is the output elasticity of capital, and ϕ measures productivity (returns) in the production of ideas (See Boix & Soler, 2017, and Boix, De Miguel and Rausell, 2021, for more detail about the derivation of the model).

3.2. Data

Boix, De Miguel and Rausell (2021) applies this model to countries, regions and municipalities. Their database comes from a wide selection of sources for each level (e.g. The World Bank, Penn World Tables, FMI, Eurostat, etc.). In this paper we use the database from Boix, De Miguel and Rausell (2021) to calculate the results using the different methods of ML.

4. Results

The results are in table 1. In methods that allows internal adjustments (e.g.: polynomial regression, XgBoost, neural networks, LLLS), the internal parameters have been fitted to produce the most accurate prediction and avoid overfitting. The interpretation of the performance/accuracy is based on the root mean squared error (RMSE) which allows a certain comparability between all the estimates.

All the models show a very high predictive capacity. The predictive performance (RMSE) of the models also depends on the sample. In the sample of 78 countries, LLLS, causal forests, polynomial regressions and random forests outperforms the other methods. In the sample of 275 European regions, the best predictive capacity comes

from one of the neural networks, LLLS, causal forests, and polynomial regression. Traditional methods such as OLS, FE and IV show a good performance, and are placed in the intermediated positions of the table. In part, this is due to the fact that the structure of the model is linear although they are penalized because most of the covariates show nonlinear patterns (e.g. the percentage of CCIs). Curiously, methods like XgBoost, regression trees and one of the neural networks do not outperform, in this case, the performance of more traditional methods.

In combined terms of performance and interpretability, LLLS and causal forest seem the most suitable, at least in the type of experiment we have performed. Both have the advantage that allows causal interpretations under unconfounding or through instruments. The polynomial or spline regressions, although they are a simple alternative with a good performance-interpretability relationship, seem to produce more unstable coefficients of the ICC effect.

Table 1. Results for the two samples. Ordered from lower to higher RMSE.

78 COUNTRIES

Method	RMSE	R2	Elasticity	SE	p-value	Impacts	Other
LLLS	0.1070	0.9853	0.1882	0.0591	0.0072	Local and global	
Causal Forest	0.1224		0.1376	0.1814	0.4507	Local and global	
GAMS/Polynomial	0.1395	0.9750	0.3516	0.1147	0.0031	Global	
Random Forest	0.1486	0.8642	-	-	-	Local and global. Using post-hoc explainers	VIP
Fixed Effects	0.2189	0.9280	0.1364	0.0561	0.0177	Global	
OLS	0.2286	0.9316	0.1595	0.0539	0.0422	Global	
Instrumental Variables	0.2339	0.9285	0.2567	0.1809	0.1605	Global	
XGBOOST	0.2479	0.9304	-	-	-	Local and global. Using post-hoc explainers	VIP
NNetwork 2*	0.2924	-	-	-	-	Local and global. Using post-hoc explainers	VIP
Conditional Tree	0.3385	0.8502	-	-	-	Local and global	
NNetwork 1*	0.5399	-	-	-	-	Local and global. Using post-hoc explainers	VIP

275 EU REGIONS

Method	RMSE	R2	Elasticity	SE	p-value	Impacts	Other
NNetwork 2*	0.0263	-	-	-	-	Local and global. Using post-hoc explainers	VIP
LLLS	0.0412	0.9712	0.1716	0.0052	0.0000	Local and global	
Causal Forest	0.0535	0.8579	0.2545	0.0295	0.0000	Local and global	
Random Forest	0.0636	0.8579	-	-	-	Local and global. Using post-hoc explainers	
GAMS/Polynomial	0.1097	0.9140	0.1230	0.0237	0.0000	Global	VIP
Fixed Effects	0.1287	0.8921	0.2469	0.0198	0.0000	Global	
OLS	0.1303	0.8894	0.2499	0.0199	0.0000	Global	
Instrumental Variables	0.1427	0.8673	0.1042	0.1215	0.3920	Global	

XGBOOST	0.1439	0.8717	-	-	-	Local and global. Using post-hoc explainers	VIP
NNetwork 1*	0.1453	-	-	-	-	Local and global. Using post-hoc explainers	VIP
Conditional Tree	0.1714	0.8086	-	-	-	Local and global	

RMSE: root mean squared error; SE: squared error; VIP: variable importance for the prediction.
* Backpropagation using only one entry layer with all the covariates, a hidden layer with 5 neurons, and an output layer with only one input.

5. Conclusions

In this paper, we depart from the framework and data used by Boix & Soler (2017) and Boix, De Miguel and Rausell (2021) for the measurement of the impacts of the CCIs on the GDP per capita and labour productivity of countries and regions and use this framework to test the performance of different machine learning methods, the differences that they produce on the estimated impacts, and the usability of those methods to evaluated the impact of the CCIs.

This framework is characterized by a linear structural equation based on a generative model that we know a priori that it will produce a good prediction, with an small number of covariates (seven) and with an small number of observations (78 and 275 in each dataset).

In this context, most of the new ML methods do not show an extraordinary advantage over traditional methods, although some of them show qualitative characteristics that do improve those of traditional methods, such as not depending on the (often unreal) assumption of linearity, directly manage heterogeneity, or produce not only global interpretations of the results (by variable), but also local interpretations (observation by observation for each variable).

In fact, some of the new ML methods, such as XgBoost, conditional trees and neural networks, have shown worse predictive ability than traditional methods. In addition, many of these new methods do not allow a direct interpretation of the impact of the ICCs, which should be done a posteriori using specific or agnostic methods, such as simplification, methods based on conditional expectations (partial dependence plots, ICE), feature importance (eg SHAP based methods), or local explanations (anchors, LIME or counterfactuals). All these post-hoc explanation methods have their limitations.

In part, the lower performance (although in no case bad) of some of these new machine learning methods has been due to the fact that their greater potential is

expressed in problems with a huge number of explanatory variables (tens, even hundreds or thousands), with very large samples (thousands or even millions of observations) and also in the absence of a defined theoretical model.

Despite this fact, it is necessary to highlight that, in the sample of regions, a simple neural network with a hidden layer of five neurons has shown a higher prediction capacity than the other methods.

The limitations of this article and its reflections are identical to future lines of research. In the first place, this article has not used all the ML techniques available, but the most representative ones, so that some others of interesting application to this context could be incorporated in the future. Secondly, it would be interesting to contrast the results with those of similar problems on the impact of ICCs, but that imply a greater number of explanatory variables, many more observations, and perhaps less structured theoretical approaches.

References

Anantrasirichai N & Bull D (2021). “Artificial intelligence in the creative industries: a review”, Artificial Intelligence Review, https://doi.org/10.1007/s10462-021-10039-7.

Boix R, De Miguel B, Rausell P (2021). “The impact of cultural and creative industries on the wealth of countries, regions and municipalities", European Planning Studies, DOI: 10.1080/09654313.2021.1909540.

Boix R, Peiró J, Rausell P (2021). “Creative industries and productivity in the European regions. Is there a Mediterranean effect?”, Regional Science Policy & Practice. DOI: 10.1111/rsp3.12395.

Boix R, Soler V (2017). “Creative service industries and regional productivity”, Papers in Regional Science, 96(2), p.261-279.

Lazzeretti L (2021). L’ascesa della società algoritmica ed il ruolo strategico della cultura. Milano: Franco Angeli.